In: SFB 100 "Elektronische Sprachforschung" (Hrsg., 1972): Aspekte der automatischen Lemmatisierung. Bericht 10-72. Linguistische Arbeiten (LA) 12, 62-81

Harald H. Zimmermann:

Das Lexikonsystem zur maschinellen Sprachbearbeitung

Wenn zur (theoretischen) Frage einer linguistischen Fundierung der Lexikonkonzeption das Problem hinzukommt, ein reales (hier: maschinelles) Wörterbuch (des Deutschen) herzustellen, ergibt sich fast notwendig ein Spannungsverhältnis zwischen den linguistischen und den technischen Gegebenheiten, oder vielleicht besser: den linguistischen Vorstellungen und den technischen Realisierungsmöglichkeiten. Man sollte meinen, dass diese Differenz eigentlich gar nicht so groß sein könne, da die moderne Linguistik mit dem Prinzip der Einfachheit einer Grammatik (u.a. der Redundanzfreiheit) gerade der Computerlinguistik entgegenkomme.

Dies mag im Hinblick auf die strukturelle Beschreibung (auch der Lexikoneinträge) durchaus zutreffen, gilt aber sicherlich nicht in vollem Umfang in der angewandten Linguistik, also auch nicht bei einem Verfahren, dessen Ziel es ist, deutsche Texte maschinell syntaktisch-morphologisch(-...) zu analysieren.

Hier spielen vor allem statistische Überlegungen eine Rolle, etwa mit dem Resultat, dass spezifische Informationen von sehr häufig auftretenden Wortformen anhand von Wortformeneinträgen ermittelt werden sollten und nicht etwa über den Wortstamm anhand des zeitraubenden 'Umwegs' einer komplizierten Wortzerlegung (Flexionsanalyse, längere Wörterbuchsuche, ...). Dabei ist der Herstellungsaufwand zu berücksichtigen: Ein (zumindest im Deutschen) bedeutend weniger umfangreiches Wortstammlexikon (mit relativ einfach zu markierenden Flexionstypenmerkmalen) leistet bei einer Textanalyse dasselbe wie ein Wortformenlexikon, bei dem die idiosynkratischen Merkmale (u.a. über eine aufwendigere Voranalyse und Notation) den Lexikoneinträgen bereits zugeordnet sind. Den Umfang des Lexikons weiter einschränken (technisch gesehen: zur Entlastung des Datenspeichers beitragen - damit ist fast stets eine kürzere Suchphase verbunden) könnte der linguistische 'Idealfall' eines Morphemlexikons, verknüpft mit einem Regelsystem, das es erlaubt, Wörter und Wortformen anhand ihrer morphologischen Bestandteile (Präfixe, Suffixe, Kernmorpheme; freie, gebundene, blockierte Morpheme) zu erkennen und zu erklären - gegenwärtig aber noch ein Wunschtraum.

Der Aufbau eines maschinellen Lexikons ist zugleich von einer Anzahl rein technischer Gegebenheiten abhängig: Hier sind vor allem Kernspeichergröße sowie Kapazitäten und Zugriffszeiten der peripheren Geräte (Magnetband/Platte/...) des verfügbaren Computers zu nennen. Die Speicherungs- und Zugriffsmöglichkeiten beeinflussen vor allem die Art der Wörterbuchsuche und damit den äußeren Aufbau des Lexikons (Hochfrequenzwörterbuch, alphabetische Ordnung der Einträge, Wortformen- bzw. Stammformenlexikon...). Daneben spielt die Frage der Datensicherung und einer möglichst in allen Phasen überschaubaren Manipulierbarkeit der Daten (etwa zur Informationskontrolle, zu Korrektur oder Ergänzung) eine Rolle.

Das Abwägen einer adäquaten linguistischen Darstellung gegen die bestmögliche (d.h. auch den gegenwärtig - etwa in Saarbrücken - vorhandenen Möglichkeiten angepasste) technische Realisierbarkeit ist dadurch erschwert, dass weder die linguistischen noch die technischen Variablen alle exakt (oder 'objektiv') definiert werden können. Man muss vielmehr davon ausgehen, dass die linguistische Konzeption eine mögliche Verifizierung eines theoretischen Modells (neben anderen möglichen Modellen) darstellt, dass aber auch die technische Konzeption in vielen Fällen nicht auf exakten Messungen oder Berechnungen aufbaut, sondern eher ein auf Erfahrung beruhendes intuitives Know-How darstellt.

Ein beredtes Zeugnis dieser Spannung (auch innerhalb der Saarbrücker Arbeitsgruppe) stellt die Behandlung der morphologisch-graphematischen Komponente des Lexikoneintrags dar. Ein erstes Beispiel dafür ist die Behandlung der Starken und Unregelmäßigen Verben des Deutschen: Noch 1970 hatte R.Dietrich die Verben (ohne allerdings so weit zu gehen wie Pike [1)] oder Billmeier [2)]) eine ausführliche formale Typisierung der Starken Verben auf der Basis der variablen und invariablen Elemente geliefert [3)] . Im maschinellen Lexikon dagegen werden die einzelnen Verbstämme (weitgehend nach traditionellem Muster: Präsens - Imperfekt - Partizip..., z.B.: GEH - GING - (GE-)GANG), abzüglich allein der Endungsflexion und des evtl. GE-Präfix beim Partizip II, wieder eingeführt; ein Stammnummernkode erlaubt es allerdings, eine Stammalternante jeweils nur einmal anzugeben. [4)]

1) Pike: Non-linear Order and anti-redundances in German Morphological Matrices, in: Zeitschr. f. dt. Mundartforschung 32(1965) S.193-221; dort vor allem die extreme 'Analyse' von SEIN (S.195).
2) Billmeier: Simulation verbalen Verhaltens, in: Vorträge der GI-Fachtagung 'Information Retrieval Systeme...', 9.-11.Dez.1970, Stuttgart, S.102-106.
3). Dietrich: Eine formale Beschreibung der Starken und Unregelmäßigen Verben der deutschen Gegenwartssprache, Arbeitsber.Nr.9 der Ling. Arbeiten des Germanist. Instituts. der Univ. d. Saarlandes (1970)
4) Vgl. LA 10, S. 64 ff

Die Behandlung des Umlauts zeigt dieses Zusammenspiel zwischen linguistischer Konzeption und technischer Ausführung noch deutlicher: Die Beschreibung der Plural-Bildung unter Abwandlung eines Stammvokals (Umlaut) bei einer Reihe von Substantiven bildet - linguistisch gesehen - kein großes Problem. So lässt sich etwa neben der Markierung der Endungsflexion ein weiteres Merkmal [+/- Umlaut] einführen, u.U. mit Angabe der Vokalstelle, das es erlaubt, die entsprechenden Wortformen zu erfassen; z.B. lässt sich aus MANN [PLURAL = ER, UMLAUT = 2, ...,] die Flexionsform MAENNER recht einfach erzeugen. Eine entsprechende Version liegt der Lochkarten-Konzeption des Saarbrücker Lexikons zugrunde (Vgl. LA 10, S.72ff.) Bei der maschinellen Identifikation einer Wortform wie MAENNER ist zur Zuordnung zum Lexikoneintrag (MANN) nicht nur eine Endungsanalyse durchzuführen (Abstreichen und Kontrolle von -ER), sondern auch die Binnenflexion zu ermitteln. Es lassen sich mehrere Erkennungsprozeduren denken, von denen die wichtigsten kurz skizziert werden sollen:

a) Als Lexikoneintrag ist - wie oben - nur MANN mit den entsprechenden Angaben vorhanden: Ein Abstreichen der Endung (Rest =MAENN) führt noch nicht zur Identifikation: es wird daher geprüft, ob ein Umlaut möglich ist; das die Zuordnung störende "E" wird entfernt. MANN wird anschließend im Lexikon gefunden; eine Kontrolle ergibt die Zulässigkeit der Endung und des Umlauts: die Identifikation war also korrekt.

b) Als Lexikoneinträge sind alle Wortformen (MANN - MANNE - MANNES - MAENNER - MAENNERN) vorgegeben: Die Wörterbuchsuche führt unmittelbar zum Eintrag MAENNER; dort ist u.a. notiert, dass es sich um eine Pluralform des Substantivs MANN handelt.

c) Als Lexikoneinträge sind MANN (wie in (a)) und MAENN (mit Hinweis auf MANN) vorgegeben: Nach Abstreichen möglicher Endungen wird der Rest MAENN mit Hinweis auf MANN gefunden; dort befinden sich schließlich die zur weiteren Identifikation nötigen Angaben.

d) Im Lexikon steht nur die Konsonantenfolge MNN (mit Hinweis auf die ausgeschnittenen Vokale, den Umlaut usw.): Vor der Wörterbuchsuche werden auch bei der zu identifizierenden Flexionsform die Vokale entfernt (MNNR); eine Zuordnung ist nach Abstreichen des Endungskonsonanten und nach entsprechender Kontrolle jetzt ebenfalls möglich. 1)

1) Ein ähnliches Verfahren bildet die Grundlage der von G.Wenzel vorgeschlagenen Lexikonstruktur in: Structure of a Lexicon in Natural Language Processing. Report Nr. 69.05.002 des Wiss. Zentrums Heidelberg der IBM Deutschland.

Grundsätzlich könnten alle diese Versionen technisch realisiert werden. Der linguistischen Basis am nächsten kommt bei diesen vier Beispielen die Version (a). Die Umlauterkennungsprozedur führt bei der Wörterbuchsuche zwar zu einigen zeitlichen Verzögerungen, das Lexikon kann aber relativ klein gehalten werden. Die Version (b) verlangt gegenüber allen anderen Versionen ein unverhältnismäßig aufgebauschtes Lexikon; da Umfang und Suchzeiten in der Regel direkt proportional sind, ist diese Lösung ökonomisch nicht besonders interessant, da der Vorteil des direkten Auffindens durch längere Suchzeiten wieder abgebaut wird. Die Version (c) hält den Umfang des Lexikons in Grenzen (vor allem, wenn man als Basis die umlautenden Simplicia ansetzt) und ermöglicht dennoch eine wenig komplexe Erkennungsprozedur, da die Suche und das Ausschneiden des (potentiellen) Umlaut-E entfallen. Die vierte Version (d) verlangt das generelle Entfernen der Vokale und ein Markieren dieser Stellen im Lexikon und in den Suchwörtern: der Lexikonumfang ist zwar stark eingeschränkt, von Nachteil ist, dass die Einträge in dieser Form nur noch schwer lesbar sind und menschliche Kontrollen erschwert werden; außerdem verlangt diese Version weitere komplexe Erkennungsregeln.

Bei der Ausdehnung dieser Verfahren auf die Erfassung der unregelmäßigen Verben - bei denen teilweise auch konsonantische Veränderungen zu beobachten sind: ZIEH - ZOG u.a.m. - versagt dieses System schließlich wieder.

Fasst man die Vorüberlegungen und die Ergebnisse des Beispiels zusammen, so zeigt sich, dass für eine Reduktion von Textwörtern auf einen Lexikoneintrag die sprachlich adäquate Beschreibung nicht ausreicht, sondern mit der technischen Realisierung in Einklang gebracht werden muss.

Dies hatte in diesem konkreten Fall zur Folge, dass die ursprüngliche Version des Saarbrücker Lexikons (Typ a) für die Wörterbuchsuche (automatisch) in die Form 'Typ c' gebracht wurde. Das Saarbrücker Lexikonsystem, dessen praktische Ausführung im folgenden näher beschrieben werden soll, ist als ein Versuch zu bewerten, beide Komponenten sinnvoll zu verbinden.

Übersicht über das Lexikonsystem

Den verschiedenartigen Problemen entsprechend wird kein einheitlich strukturiertes Lexikon zugrunde gelegt; vielmehr handelt es sich um separate Lexika, die zu einem ineinanderwirkenden System zusammengefügt sind. Im Mittelpunkt steht dabei das so genannte BASIS-LEXIKON, in dem alle Lemmata mit den zugehörigen Informationen aufgeführt sind. Daneben ist - ein in der maschinellen Sprachbearbeitung bewährtes Verfahren - ein HOCHFREQUENZWOERTERBUCH zur raschen Erfassung der häufigsten Wortformen bereitgestellt. Ein Sonderlexikon STVRB steht zur Erkennung der unregelmäßigen Verben zur Verfügung. Ergänzt wird dieses System durch ein LEXIKON DER FESTEN SYNTAGMEN. Präfix- und Suffixlisten sollen darüber hinaus erlauben, Wortklassifizierungen durchzuführen, falls die Lexikoneintragungen dafür nicht ausreichen. Die einzelnen Lexika werden (mit Ausnahme des Lexikons der festen Wendungen [1)] in Aufbau und Funktion im folgenden kurz beschrieben

1) Vgl. dazu den Artikel von A.Rothkegel in diesem Bericht.

Das Basislexikon

Wie bereits in LA 1O, S.86ff. angedeutet wurde, bildet die Lochkarte den primären Datenträger des Lexikons zur Lemmatisierung. Aus Gründen der Datensicherung werden alle Korrekturen und

Ergänzungen auch auf diesen Ausgangsdaten notiert. Die Lochkarten enthalten also zunächst alle Angaben und Markierungen, die in den Kodierungsanweisungen in LA 1O beschrieben sind. Ein getreues Bild der Lochkarten wird (mittels des Programms LKBILD) - u.a. zur Datensicherung - auf Magnetband übertragen; davon ausgehend wird ein weiteres Lexikon (mittels des Programms LEMMWOBU) bereitgestellt, das die Daten zur vereinfachten Kontrolle in umkodierter Form enthält. Es steht ebenfalls auf Magnetband zu einer weiteren maschinellen Verarbeitung zur Verfügung. Bei der Herstellung dieses Lexikons - es verlangt eine alphabetische Ordnung der Einträge - werden Alphabetkontrollen durchgeführt (d.h., dass die Lochkarten bereits alphabetisch vorgeordnet sein müssen) und formal unzulässige Kodierungen festgestellt. Eine wesentliche Aufgabe des Programms LEMMWOBU ist es darüber hinaus, die bei der Erstellung der morphologischen Einträge nicht ausdrücklich als Präfixe oder Suffixe markierten Morpheme anhand entsprechender Listen zu klassifizieren (Nur bei Ausnahmen - Mehrdeutigkeiten - sind bei der Aufnahme Zusatzangaben mitgegeben, um falsche Einordnungen auszuschließen). Je Lexikoneintrag ist im Basislexikon gegenwärtig eine feste Länge von 42 Zellen (zu 24 Bits) als Grundlage der Informationsspeicherung gewählt.

Übersicht zur Umkodierung Lochkartenbild - Basislexikon (Band)

A Allgemeiner Teil [1)]

Speicherzelle [4)]	Inhalt
1 - 8	Worteintrag I (ohne Sonderzeichen)
9 - 12	Worteintrag II (z.B. Verweis auf Lemma)
13	Wortklassenangabe ("A","S" oder "V" rechtsbündig)
14+15	Bei Kompositum: Hinweis auf weiteres Lernma [2)]
16-18	Grenzenmarkierung [3)]
19+20	Präfixe (= Präfixe. der Grenze, ab der Präfix beginnt)
21+22	Stamm -Morphem (= Nr. der Grenze, ab der Stamm-Morphem beginnt)
23+24	Suffixe (= Nr. der Grenze, ab der Suffix beginnt)
25+26	Fugen (=Nr. der Grenze, ab der Fuge beginnt)
41+42	Kennzeichnung von "ß" (Stelle und Art)

1) Bei Adjektiven (A), Verben (V) und Substantiven (S) gleich
2) Auf der Lochkarte durch "/" markiert; hier: Nr. der "Grenze"
3) "Grenze" = Graphem-Nr. im Wort, mit dem ein weiteres Morphem (Stamm, Fuge, Suffix, Präfix) beginnt. Das erste Morphem beginnt stets mit dem ersten Graphem und entfällt bei der Notierung; die letzte Notierung bezieht sich auf das "Blank" (Wortende).
4) Die Angaben in den Zellen 14-26 stehen linksbündig in Character-Notation in einer Zelle. Eine Zelle fasst maximal 4 Werte (je 6 Bits), z.b. - " | 1 | 3 | 4 | 9 |

B Adjektive

SP [1)]	LK [2)]	Abkürzung	Bedeutung
27/1	30	IUL	Umlautkennzeichen
/2	32	IUN	Negationspräfigierung
/3	33	ISYV	Syntaktische Verwendung
/4	34	IST	Steigerung
28/1	40	IUW	Unregelmäßigkeit bei Worteintrag I
29/1	42	KMP	Komparativ-Einschränkung
30/1	39	NS	Nebensatzrektion
/2	80	ZW	Zweifelsfall

/3	41	GR	Graphemvariante / NF
/4	-	$=1	Sonderzeichen im Wortlaut
31	31	KL	Morphologische Klassifizierung
32	35-38	IREKT	Präpositionale Rektion
33/4	43	IE	Eigenschaft-Subkategorisierung

1)Speicherzellen-Nr. im Informationsfeld, ggf. nach dem Schrägstrich die Charakter-Nr. innerhalb eines Speicherwertes
2) Lochkartenspalte, in der die entsprechende Information bei der Eingabe vermerkt war

Verben

SP	LK [3]	Abkürzung	Bedeutung
27/1	30	BED	Bedeutungsnummer
/2	33	IGE	Gemischte Endungsflektion
/3	34	IE	Infinitivendung
/4	37	IST	Steigerung Partizipien
28/1	38	IZ1	Zusatzinformation I
/2	39	IZ2	Zusatzinf. II (Modalverben)
/3	40	IZ3	Zusatzinf. III (Perfektbildung) + "sein"
/4	41	IREF	Reflexivität
29/1	35	IFLE	Eingeschränkte Flektierbarkeit

3) Eine nachgestellte II zeigt an, daß die Information der zweiten Verbkarte entnommen ist.

SP	LK	Abkürzung	Bedeutung
29/2	43	IAPR	Abtrennbare Präfixe
/3	42	IPR	Präfixmorpheme
30/1	38 (II)	NS	Nebensatzrektion
/2	80	ZW	Zweifelsfall
/3	36	IFLP	Flektierbarkeit der Partizipien
/4	-	$=1	Sonderzeichen im Wortlaut
31	31+32	ISTK	Stammnummernkode
32	39-41(II)	IREKT	Präpositionale Rektion
33	42-47(II)	IBL1	Selektionale Subkategorisierung
34	48-53(II)	IBL2	"
35	54-59(II)	IBL3	"
36	60-65(II)	IBL4	"
37	66-71 (II)	IBL5	"
38	72-77(II)	IBL6	"
39/1	31(II)	ISRE1	Strikte Subkategorisierung (Rektion)
/2	32(II)	ISRE2	"
/3	33(II)	ISRE3	"
/4	34(II)	ISRE4	"
40/1	35(II)	ISRE5	"
/2	36(II)	ISRE6	"
/3	37(II)	ISRE7	"

Substantive

SP	LK	Abkürzung	Bedeutung
27/1	39	IUL	Umlautangabe
/2	38	INUM	Numerusbeschränkungen
/3	40	IFLG1	Flexion im Genitiv Singular (I)
/4	41	IFLG2	" (II)
28/1	44	IS1K	Sonderformtyp 1
/2	50	IS2K	" 2
/3	56	IS3K	" 3
/4	62	IS4K	" 4

SP	LK	Abkürzung	Bedeutung	
29/1	45	IS1A	Sonderformgraphenzahl 1	
/2	51	IS2A	" 2	
/3	57	IS3A	" 3	
/4	63	-		
30/1	68	NS	Nebensatzrektion	
/2	80	ZW	Zweifelsfall	
/3	78	NF	Nebenform	
/40	-	$=1	Sonderzeichen im Wortlaut	
31	37	IGEN	Genusangabe	
32	69+70	IPRP	Präpositionale Rektion	
33	46-49	IS1G	Wortlaut d. unregelm. Flexionsmorphems	1
34	52-55	IS2G	"	2
35	58-61	IS3G	"	3
34	64-67			
37	71-75	ISK	Subkategorisierung (Bits)	
38	42	IFLN1	Flexion im Nominativ Plural (I)	
39	43	IFLN2	" (II)	
40	-	-	Singularendung bei unregelm. Flexion	
	43-46	IBED	Bed.-Nr. der Morpheme + Stelle	

Die Magnetbandversion des Basislexikons ist zunächst die Grundlage statistischer Untersuchungen, die gegenwärtig vor allem zur Lexikonkorrektur verwendet werden. Ausgehend von den Informationen des Basislexikons werden dabei die verschiedenartigsten Prüflisten erstellt: vorwiegend handelt es sich dabei um 'inverted files', d.h. Ordnungen (und Aufzählungen) der Einträge nach bestimmten Merkmalen (etwa bei den Verben: nach der Art der Perfekt-Bildung, d.h. Gruppierung aller Verben, die das Perfekt mit SEIN, mit HABEN oder mit SEIN v HABEN bilden). Für die ca. 12.000 Adjektive und die etwa 16.000 Verben, die bereits anhand des 'Wahrig' vollständig erfasst sind, liegen entsprechende Prüflisten und statistische Auswertungen bereits vor; rund 15.000 Substantive (Anfangsbuchstaben A - E) sind ebenfalls in der ersten Prüfphase. Darüber hinaus werden anhand bestimmter (vor allem die Flexion betreffender) Merkmale Wortformenlisten generiert, die die Überprüfung der Markierungen erleichtern. Beispielsweise wird anhand der Kennzeichen bei Substantiven die für die Flexion typischen Wortformen Genitiv-Singular und Nominativ-Plural erzeugt, bei HAUS also HAUSES und HAEUSER; bei Verben - in der Regel aus dem Infinitiv-Eintrag - die 3. Person Singular Indikativ Aktiv (vor allem zur Kontrolle der Abtrennbarkeit von Präfixen/Verbzusätzen) und das Partizip II (zur Kontrolle des GE-Partikels): Also bei dem Eintrag AUSSAG die Formen SAGTE AUS und AUSGESAGT. Diese Auswahl und Generierung übernimmt das Programm AUSWLEMM. Zur Überprüfung der Adjektive steht bereits ein komplexeres

Generierungssystem zur Verfügung (Programm ADJGEN), bei dem neben Steigerungsformen und -möglichkeiten vor allem kontextbezogene Merkmale (DASS-Anschluss oder etwa die attributive Verwendbarkeit anhand bestimmter semantischer Substantiv-Subklassen) einbezogen sind.

Die Magnetband-Version des Basislexikons ist für die Wörterbuchsuche aufgrund bedeutend langsamerer Zugriffseinheiten wenig geeignet: Diese Version ist daher nur eine Zwischenstufe zu einer kompakten Magnetplattenversion und bildet die Grundlage für die zur Komprimierung notwendigen statistischen Ermittlungen. Funktion und Aufbau des kompakten Lexikons beschreibt der Artikel von P. Krebs in diesem Bericht.

Das Hochfrequenzwörterbuch (HFWB)

Es ist bekannt, dass eine geringe Zahl von Wortformen (als types) bei einem fortlaufenden Text einen hohen Anteil (als tokens) ausmachen. Diese Erkenntnis wird bei der Wörterbuchsuche ausgenutzt. Ein Häufigkeits-Wörterbuch, orientiert an eigenen Zählungen (ca. 200.000 Wortformen wissenschaftlicher Prosa und Zeitungstexte [1)] und an Meiers Auswertung des Kaeding [2)] ergab eine Liste von 127 Wortformen, die bei der Wörterbuchsuche ständig im Kernspeicher verfügbar ist und mittels eines binären Suchverfahrens einen Textanteil von ca. 40 - 50 % der laufenden Wortformen erfassen soll: vgl. Liste (unten), Die entsprechenden grammatischen Merkmale dieser (teilweise mehrdeutigen) Wortformen können über eine weitere Liste zugeordnet werden und sind bei einem positiven Such-Ergebnis anzufügen.

1) Vgl. dazu Eggers, H.: Zur Syntax der deutschen Sprache der Gegenwart, In: Stud. gen. 15, 1962, S. 49-59.
2) Siehe Meier,H.: Deutsche Sprachstatistik, Hildesheim 1964.

WORTLISTE HÄUFIGKEITSWÖRTERBUCH

ABER ALLE ALS ALSO AM AN ANDERE ANDEREN AUCH AUF AUS BEI BEIDEN BIS DA DAMIT DANN DAS DASS DEM DEN DENN DER DES DIE DIESE DIESEM DIESEN DIESER DIESES DOCH DURCH EIN EINE EINEM EINEN EINER EINES EINMAL ER ERSTEN ES ETWAS FUER GANZ GEGEN GIBT GROSSEN HABE HABEN HAT HATTE HIER IHM IHN IHR IHRE IHREN IHRER IM IMMER IN IST JAHRE JETZT KANN KEIN KEINE KOENNEN MACHT MAN MEHR MIT MUESSEN MUSS NACH NICHT NICHTS NOCH NUN NUR OB ODER OHNE RECHT SCHON SEI SEIN SEINE SEINEN SEINER SELBST SICH SIE SIND SO SOLL SONDERN UEBER UM UND UNS UNTER VIELE VOM VON VOR WAERE WAR WAREN WAS WEIL WELT WENN WERDEN WIE WIEDER WIR WIRD WORDEN WURDE WURDEN ZEIT ZU ZUM ZUR ZWEI

Das Sonderlexikon der unregelmäßigen Verben (STVRB)

Die Berücksichtigung der unregelmäßigen Verben stellte bei der Errichtung des Basislexikons ein besonderes Problem dar. Es gibt im Deutschen nahezu 200 Simplizia, bei denen - abgesehen von den Endungen - graphematische Abweichungen des 'Stammes' bei der Flexion zu beobachten sind ; - von Extremfällen wie IST und WAR (zu SEIN), bei denen keine gemeinsamen Grapheme mehr vorhanden sind, über starke Modifikationen (wie ZIEH - ZOG, BRING - BRACHT) bis zu den 'einfach' ablautenden SING - SANG - SAENG - SUNG.

Die kaum abzusehende Zahl der aktuellen oder potentiellen Präfigierungen ließ eine Aufnahme aller dieser Einträge einschließlich der Präfixe in das Basislexikon nicht sinnvoll erscheinen. Da es dennoch nötig war, die präfigierten Verben mit den ihnen idiosynkratisch zukommenden, also über Ableitungsregeln nicht fassbaren grammatischen Informationen aufzunehmen - LAUFEN etwa

verlangt andere Subklassifizierungen als VERLAUFEN und WEGLAUFEN - wurde der Infinitiv-Stamm aller Verben einschließlich der evtl. vorhandenen Präfixe im BASISLEXIKON verzeichnet.

Die Zuordnung einer Wortform wie LIEF oder ERKLANG erfolgt dabei mithilfe des Speziallexikons, das alle abweichenden, unregelmäßigen "Stämme" der Verb-Simplizia unter Hinweis auf den potentiellen Infinitiv-Stamm enthält. So verweisen SANG, SAENG, SUNG auf SING, LIEF auf LAUF oder ZOG, ZOEG auf ZIEH. Ein Kodesystem regelt darüber hinaus den Anschluss der korrekten Endungen. Um den generellen Wortbildungsmöglichkeiten Rechnung zu tragen, werden die Präfixe (oder Verbzusätze) nicht berücksichtigt, d.h. eine Wortform wie KOPFSTANDEN wird ebenso behandelt wie WEGLIEF (=> WEGLAUF). Einige kontextsensitive graphematische Regeln ordnen die korrekte Zuordnung von Wortformen wie ABRIET (nicht fälschlich A-BRIET, sondern AB-RIET =>ABRAT), ZUSAMMENBRIET (nicht ZUSAMMENB-RIET, sondern ZUSAMMEN-BRIET => Z.BRAT), ABFRISS (AB-FRISS => ABFRESS), AUFRISS (nicht AU-FRISS, sondern AUF-RISS => AUFREISS) oder HERUNTERISST (HERUNTER-ISST => H.ESS).

Einige Probleme der Lexikoneinträge, die sich im Zusammenhang mit der Wörterbuchsuche ergeben, seien abschliessend noch erörtert:

Nicht alle Flexionsformen der Substantive lassen sich regularisieren. Man könnte daher vorsehen, bei unregelmäßig flektierenden Substantiven künstliche Stämme im Lexikon zu bewahren, die die gemeinsamen Grapheme eines 'Lemmas' darstellen. So müssten abweichend von der Singularform etwa KAKTUS, PRONOMEN, MUSEUM zu KAKT, PRONOM, MUSE verkürzt werden, um zu gewährleisten, dass ohne zusätzliche Maßnahmen die Plural-Endungen zugeordnet werden können; KAKTEEN, PRONOMINA und MUSEEN könnten dann über diesen einen Eintrag identifiziert werden. Hier sind jedoch zugunsten eines einfacheren Erkennungsalgorithmus zwei Einträge (Plural- und Singular-) vorgesehen. Umlautende Adjektive und Substantive (z.B. AELTER, MAENNER) werden ebenfalls im Basislexikon über die Morphemalternanten (z.B. AELT mit Hinweis auf ALT, MAENN mit Verweis auf MANN) identifiziert. Auf diese Weise soll die zeitraubende Untersuchung der Binnenflexion weitgehend ausgeschlossen werden. Damit ist einer linguistisch adäquaten Behandlung dieses Problems nicht vorgegriffen: bei potentiellen Wortableitungen (etwa der Bildung von Diminutiva wie HUENDCHEN aus HUND - keine Umlaut-Kennung über die Flexion möglich -) stellt sich dieses Problem in anderem Zusammenhang erneut und muss hier über Regeln bewältigt werden.

Ein Regelsystem, dessen Prinzipien an anderer Stelle beschrieben werden [1)] , hat auch die Aufgabe, nicht im Lexikon enthaltene Einträge anhand von Wortbildungskriterien (Wortableitung und Komposition) zu klassifizieren. In diesem Zusammenhang bedarf die Frage der Aufnahme von Einträgen (im Hinblick auf eine Erweiterung, aber auch eine Kürzung des Lexikons) weiterer Diskussionen. Wie bereits in LA 10, S. 16f. erwähnt, wurde bisher das Lexikon von G.Wahrig relativ kritiklos übernommen. Eine Modifikation dieses Lexikons - und damit des Saarbrücker Basislexikons - bleibt weiteren Untersuchungen vorbehalten. Zum Abschluss die Liste der Einträge des Verbsonderlexikons STVRB (454 Belege):

Die erste Rubrik UST dient zur graphematischen Identifikation des 'unregelmäßigen` Stammes' einer Verbform, dem dann (etwa zur weiteren Wörterbuchsuche) die unter der Rubrik IST stehende entsprechende 'Infinitiv-Stammform' zugeordnet wird. Die Zahlen der darauf folgenden Rubriken UK bzw. IK entsprechen den in LA 10, S.57 notierten Stammnummernkodes (binäre Addition). Die letzte Rubrik EK gibt Auskunft über die bei der UST-Form zugelassenen Endungen; falls der Stammnummernkode UST die Zahl 16 enthält (= Partizip-II-Stamm) sind darüber hinaus die adjektivischen Flexionsendungen zugelassen.

Die Zahlen von EK repräsentieren (binäre Addition) die folgenden Endgrapheme:

1 = 0	8 = ET bzw. EST
2 = E	16 = T
4 = ST	32 = EN

1) Vgl. den Artikel von Schmidt/Thiel in diesem Bericht.